Diplomski studij

Informacijska i komunikacijska tehnologija

Telekomunikacije i informatika

Računarstvo

Programsko inženjerstvo i informacijski sustavi

Računalno inženjerstvo
Računarska znanost

Raspodijeljena obrada velikih skupova podataka

2. Laboratorijska vježba

Ak. g. 2016./2017.

Zadatak 1: Korištenje obrasca za filtriranje podataka (*filtering pattern*)

Za potrebe laboratorijskih vježbi pripremljena je datoteka `sorted_data.csv` s natjecanja DEBS Grand Challenge 2015 koja sadrži podatke o vožnjama taksija na širem području grada New Yorka te je proširena infromacijom o cijeni vožnje. Na vlastito računalo preuzmite ulaznu datoteku (http://svn.tel.fer.hr/sorted_data.tar.gz) koju ćete koristiti za razvoj i debuggiranje vaših MapReduce programa. Vrijednosti u datoteci su odvojene zarezom i kreću od prvog retka (tj. datoteka ne sadrži zaglavlje), a struktura datoteke je sljedeća:

medallion	an md5sum of the identifier of the taxi - vehicle bound
hack_license	an md5sum of the identifier for the taxi license
pickup_datetime	time when the passenger(s) were picked up
dropoff_datetime	time when the passenger(s) were dropped off
trip_time_in_secs	duration of the trip
trip_distance	trip distance in miles
pickup_longitude	longitude coordinate of the pickup location
pickup_latitude	latitude coordinate of the pickup location
dropoff_longitude	longitude coordinate of the drop-off location
dropoff_latitude	latitude coordinate of the drop-off location
payment_type	the payment method - credit card or cash
fare_amount	fare amount in dollars
surcharge	surcharge in dollars
mta_tax	tax in dollars
tip_amount	tip in dollars
tolls_amount	bridge and tunnel tolls in dollars
total_amount	total paid amount in dollars

Zadatak

Cilj zadatka je napisati MapReduce program koji implementira obrazac filtriranja podataka prema definiranom kriteriju (*filtering pattern*).

Uspješnim rješenjem ovog zadatka steći ćete sljedeća znanja:

- **oblikovanje programa MapReduce prema zadanom obrascu**
- **prilagodba postavki programa MapReduce**

Detaljni opis zadatka

Proučite obrazac filtriranja podataka prema definiranom kriteriju. Osmislite rješenje koje će iz ulaznog skupa podataka odstraniti sve taksi vožnje čija je cijena 0 ili manja (parametar: *`total_amount`*) te počinju ili završavaju izvan promatranog geografskog područja. Područje je definirano mrežom ćelija (kvadrata) veličine 1000 m x 1000 m s početkom u ćeliji 1.1 koja se nalazi na geografskoj lokaciji (latitude, longitude) [41.474937, -74.913585]. Oznake ćelija povećavaju se prema istoku i jugu tako da je pomak prema istoku prva, a pomak prema jugu druga komponenta oznake ćelije (npr. ćelija 3.7 nalazi se 2 ćelije istočno i 6 ćelija južno od ćelije 1.1). Cjelokupno područje proteže se 150 km južno i 150 km istočno od ćelije 1.1, pri čemu je

ćelija 150.150 posljednja ćelija u mreži. Sve vožnje koje započinju ili završavaju izvan ovog područja tretiraju se kao odstupanja i potrebno ih je odstraniti iz ulaznog skupa podataka. Pretpostavite da pomak od 1000 metara u smjeru juga iznosi: 0.008983112, dok pomak u smjeru istoka iznosi: 0.011972, tj. posljednja ćelija 150.150 ima geografske koordinate [40,1274702, -73,117785].

Određivanje identifikatora ćelije dano je sljedećim pseudo-kodom:

```
cellId[0] = ((int)(lon - BEGIN_LON) / GRID_LENGTH) + 1
cellId[1] = ((int)(BEGIN_LAT - lat) / GRID_WIDTH) + 1
```

Program treba napraviti izlazne datoteke koje će sadržavati podatke o vožnjama koje zadovoljavaju zadane kriterije istovjetne ulaznom podatkovnom skupu.

Nakon uspješnog pokretanja odgovorite na sljedeća pitanja:

- Koje funkcije sadrži vaš MapReduce program (map, reduce, partitioner, combiner, itd.)?
- Koliko izlaznih datoteka je nastalo nakon izvođenja vašeg MapReduce programa?
- Koliko zapisa u ulaznoj datoteci nije zadovoljilo zadane kriterije?

Zadatak 2: Obrada podataka korištenjem programa MapReduce

(*NAPOMENA: kao ulaznu datoteku u zadatku koristite izlazne datoteke iz prethodnog zadatka)*

Zadatak

Cilj zadatka je napisati program MapReduce koji će na satnoj razini (neovisno o danu u tjednu) identificirati ćeliju s najviše vožnji i najvećim prihodom.

Uspješnim rješenjem ovog zadatka steći ćete sljedeća znanja:

- **definiranje programa MapReduce koji koristi složene strukture podataka**
- **prilagodba postavki programa MapReduce**

Detaljni opis zadatka

U ovom zadatku napravit ćete program MapReduce koji će obraditi podatke iz ulazne datoteke tako da odredite ćeliju u kojoj je ostvaren najveći prihod te ćeliju s najvećim brojem taksi vožnji za svaki sat u danu (neovisno o danu u tjednu). Pretpostavite da vožnja pripada samo onoj ćeliji u kojoj je započela, a prilikom računanja prihoda koristite parametar *`total_amount`*.

Program treba napraviti izlazne datoteke koje sadrže 3 retka:

- promatrano vrijeme vožnje (sat),
- oznaka ćelije s najvećim brojem vožnji i ukupan broj ostvarenih vožnji u tom satu,
- oznaka ćelije u kojoj je ostvaren najveći prihod i ukupan prihod ostvaren u tom satu.

Nakon uspješnog pokretanja odgovorite na sljedeća pitanja:

- Koje funkcije sadrži vaš MapReduce program?
- Koje tipove podataka ste definirali kao izlaz iz MapReduce programa?

Zadatak 3: Izvršavanje ulančanih MapReduce programa

(*NAPOMENA: zadatak se temelji na programima MapReduce iz zadatka 1 i 2)*

Zadatak

Cilj zadatka je napisati program koji se sastoji od dva ulančana MapReduce programa. Prvi MapReduce program priprema podatke za obradu, dok drugi MapReduce program obrađuje podatke i zapisuje ih u izlaznu datoteku.

Uspješnim rješenjem ovog zadatka steći ćete sljedeća znanja:

- **definiranje ulančanog MapReduce posla**

Detaljni opis zadatka

Ovaj zadatak se temelji na prethodna dva, tj. cilj je napraviti novi MapReduce program koji ulančava prethodna dva zadatka uz manje promjene. Program treba na satnoj razini na osnovu svih taksi vožnji čija je cijena veća od 0 te su ostvarene unutar promatranog geografskog područja identificirati ćeliju u kojoj je ostvaren najveći prihod te ćeliju s najvećim brojem taksi vožnji. Za potrebe ovog zadatka definirajte novi obrazac filtriranja koji na izlazu daje samo podatke nužne za daljnju obradu (tj. podatke koji se koriste na ulazu u drugi MapReduce zadatak).

Ulančani MapReduce program treba napraviti izlazne datoteke koje sadrže 3 retka (isto kao i u zadatku 2):

- promatrano vrijeme vožnje (sat),
- oznaka ćelije s najvećim brojem vožnji i ukupan broj ostvarenih vožnji u tom satu,
- oznaka ćelije u kojoj je ostvaren najveći prihod i ukupan prihod ostvaren u tom satu.

Na kraju programa programski obrišite međurezultate (izlaz iz svakog MapReduce posla koji nije posljednji).

Nakon što ste izvršili program u lokalnom načinu rada prebacite ga u Hadoop okružje i izvedite u grozdu Cloudera. Kao ulaznu datoteku koristite sadržaj mape `/user/rovkp/debs2015full`, a izlazne datoteke zapišite u vlastitu mapu kao i na 1. laboratorijskoj vježbi.

Nakon uspješnog pokretanja odgovorite na sljedeća pitanja:

- Kolika je veličina ulaznog skupa podataka?
- Koliko ste MapReduce poslova izvršili u vašem kôdu?
- Koje ste promjene napravili za prvi, a koje za drugi MapReduce posao?
- Koje su prednosti, a koji nedostaci ovakvog sažimanja međurezultata?